Вероятность переобучения для плотных и разреженных сеток алгоритмов

Фрей Александр Ильич

Московский физико-технический институт
(Государственный университет)
Кафедра «Интеллектуальные Системы» (ВЦ РАН)

Научный руководитель: к.ф.-м.н. Воронцов Константин Вячеславович

20 октября 2010

Проблема переобучения: комбинаторный подход

- $\mathbb{X} = (x_1 \dots x_L)$ — генеральная выборка объектов, $x_i \in \mathcal{X}$;
- $\mathbb{Y} = (y_1 \dots y_L)$ — вектор классов объектов, $y_i \in \mathcal{Y}$;
- $a : \mathcal{X} \to \mathcal{Y}$ — алгоритм классификации;
- ошибка алгоритма: $a(x_i) \neq y_i$;
- $n(a, U)$ — число ошибок алгоритма на подвыборке $U \subset \mathbb{X}$;
- $\nu(a, U) = \frac{n(a,U)}{|U|}$ — частота ошибок;
- $X^\ell \subset \mathbb{X}$ — обучающая выборка длины ℓ;
- $X^k = \mathbb{X} \backslash X^\ell$ — контрольная выборка длины $k = L - \ell$;
- разность частоты ошибок на контроле и обучении:

$$\delta(a, X^\ell) = \nu(a, X^k) - \nu(a, X^\ell);$$

- $\mu\colon \{X^\ell\} \to \mathbb{A}$ — детерминированный метод обучения;
- вероятность переобучения:

$$Q_\mu(\varepsilon) = \mathsf{P}\Big[\delta(\mu X^\ell, X^\ell) \geq \varepsilon\Big], \text{ где } \mathsf{P} \overset{def}{=} \frac{1}{C_L^\ell} \sum_{X^\ell \in [\mathbb{X}]^\ell}$$

Проблема переобучения: выбор лучшего алгоритма

- $\mu X = \underset{a \in \mathbb{A}}{\operatorname{argmin}}\, n(a, X^\ell)$ — детерминированный МЭР;
- Вероятность переобучения:

$$Q_\mu(\varepsilon) = \mathsf{P}\,[\delta(\mu X^\ell, X^\ell) \geq \varepsilon] = \mathsf{P} \sum_{a \in A} [\mu X^\ell = a][\delta(a, X^\ell) \geq \varepsilon].$$

- Рандомизированная минимизация эмпирического риска:

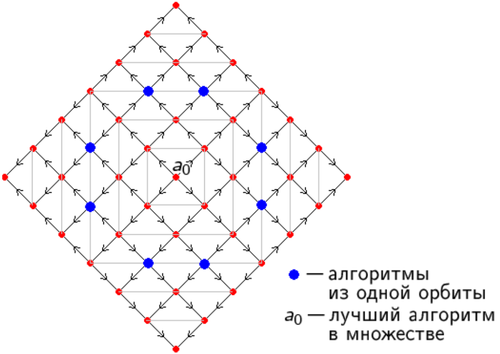
$$\mu(A, X, a) = \frac{[a \in A(X)]}{|A(X)|}, \text{ где } A(X) = \underset{a \in A}{\operatorname{Argmin}}\, n(a, X);$$

- Вероятность переобучения для рандомизированного метода обучения:

$$Q_\mu(\varepsilon, A) = \mathsf{E} \sum_{a \in A} \mu(A, X, a)[\delta(a, X) \geq \varepsilon]$$

Группа симметрий множества алгоритмов

Граф смежности двумерной унимодальной сетки:



- S_L — группа всех перестановок объектов выборки,
- S_L действует множестве всех алгоритмов $2^{\mathbb{A}}$,
- $\mathrm{Sym}(A) = \{\pi \in S_L : \pi A = A\} \subset S_L$.
- Орбита алгоритма a это $\{\pi a : \pi \in \mathrm{Sym}(A)\} \subset A$

Равный вклад алгоритмов одной орбиты

- Вероятность переобучения — сумма вкладов алгоритмов:

$$Q_\mu(\varepsilon, A) = \sum_{a \in A} Q_\mu(\varepsilon, a, A), \text{ где}$$

$$Q_\mu(\varepsilon, a, A) = \mathbf{E}\mu(A, X, a)[\delta(a, X^\ell) \geq \varepsilon];$$

- Алгоритмы одной орбиты дают равный вклад:

$$Q_\mu(\varepsilon, a, A) = Q_\mu(\varepsilon, \pi a, A), \text{ где } \pi \in \mathrm{Sym}(A)$$

- Обозначим $\Omega(A)$ — множество орбит $\mathrm{Sym}(A)$ на A;
- Вероятность переобучения с учетом структуры множества алгоритмов:

$$Q_\mu(\varepsilon, A) = \sum_{\omega \in \Omega(A)} |\omega| \, \mathbf{E}\mu(A, X, a) \left[\delta(a_\omega, X^\ell) \geq \varepsilon\right]. \qquad (1)$$

Равный вклад разбиений одной орбиты

- Вклад разбиения $X^\ell \in [\mathbb{X}]^\ell$ в вероятность переобучения РМЭР:

$$\phi(A, X, \varepsilon) = \frac{1}{|A(X)|} \sum_{a \in A(X)} [\delta(a, X) \geq \varepsilon];$$

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X^\ell \in [\mathbb{X}]^\ell} \phi(A, X, \varepsilon);$$

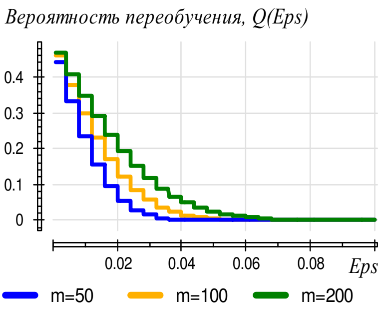
- Разбиения одной орбиты дают равный вклад:

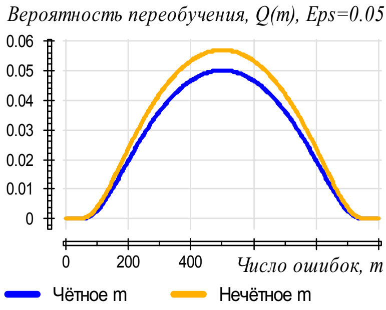
$$\phi(A, X, \varepsilon) = \phi(A, \pi X, \varepsilon), \text{ где } \pi \in \mathrm{Sym}(A);$$

- Обозначим $\Omega(\mathbb{X})$ — множество орбит $\mathrm{Sym}(A)$ на $[\mathbb{X}]^\ell$;
- Вероятность переобучения с учетом структуры множества алгоритмов:

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega(\mathbb{X})} |\tau| \phi(A, X_\tau, \varepsilon).$$

Overfitting probability for fixed predictor





Теорема (Overfitting probability for fixed predictor)

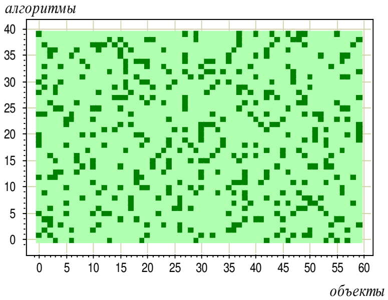
$$Q_{\mu(f)}(\varepsilon) = P\Big\{\delta_\mu(X^\ell, X^k) \geq \varepsilon\Big\} = H_L^{\ell,m}(s_0),$$

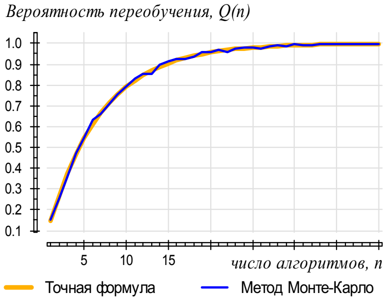
where $m = n(f, \mathbb{X})$, $s_0 = \frac{\ell}{L}(m - \varepsilon k)$, $H_L^{\ell,m}(s_0) = \sum\limits_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$.

Результаты, полученные для РМЭР

- **Связка монотонных цепочек**
  - Фрей А. И., Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 66–69.
- **Шар алгоритмов и центральный слой шара**
  - Толстихин И. О., Точная оценка вероятности переобучения для одного специального семейства алгоритмов // Конференция «Ломоносов-2010».
- **Полный слой и полный куб алгоритмов**
  - Frei A. I., Accurate Estimates of the Generalization Ability for Symmetric Sets of Predictors and Randomized Learning Algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 241-250.
- **Монотонные и унимодальные сетки**
  - Фрей А. И., Вероятность переобучения плотных и разреженных семейств многомерных сеток алгоритмов // Международ. конф. ИОИ-8 — М.: МАКС Пресс, 2010. — С. 87–90.

Overfitting probability: predictors with random errors





- A^n_m — set of n predictors, with m errors for each one. Errors are not correlated.

Теорема (Overfitting probability for A^n_m)

Let μ — randomized ERM. Then

$$\boldsymbol{E}_G P_{\mathbb{F}}(\varepsilon, A^n_m) = 1 - (1 - P_{\mathbb{F}}(\varepsilon, a_m))^n$$

Отсутствие связности: случайные множества алгоритмов

- Пронумеруем алгоритмы: $A = (a_1, \ldots, a_d)$
- S_L — симметрическая группа порядка L;
  - S_L действует на объектах выборки;
  - S_L действует на векторах ошибок алгоритмов;
- $G = (S_L)^d$ — свободное произведение S_L;
  - $g = (g_1, \ldots, g_d) \in G$ — элемент группы G, $g_i \in S_L$;
  - G действует на векторе алгоритмов:

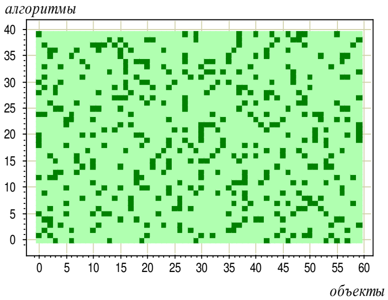
$$gA = \bigl(g_1(a_1), \ldots, g_d(a_d)\bigr).$$

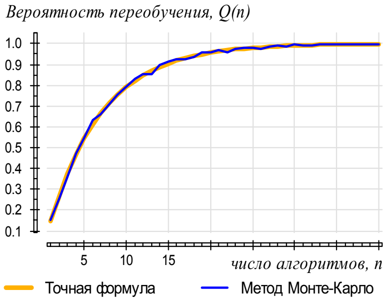
- Вероятность переобучения несвязной перестановки A:

$$\bar{Q}_\mu(\varepsilon, A) = \mathbf{E}_G\, Q_\mu(\varepsilon, gA), \text{ где } \mathbf{E}_G \stackrel{def}{=} \frac{1}{(L!)^d} \sum_{g \in G}$$

- $\bar{Q}_\mu(\varepsilon, A)$ зависит только от профиля расслоения A.

Проблема переобучения: случайный слой алгоритмов





- A_m^n — множество из n алгоритмов, допускающих по m ошибок. Ошибки расположены «случайным» образом.

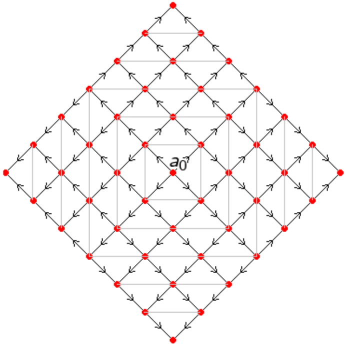
Теорема (Вероятность переобучения для A_m^n)

Пусть μ — рандомизированный МЭР. Тогда

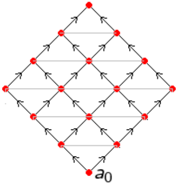
$$\mathbf{E}_G Q_\mu(\varepsilon, A_m^n) = 1 - (1 - Q_\mu(\varepsilon, a_m))^n$$

Continuous predictors set

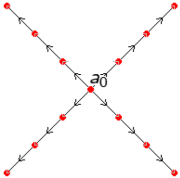
Let us study behavior of the following predictors set:



Унимодальная сетка



Монотонная сетка



Связка монотонных цепочек

- A_B — Monotonic chains binding of h, length D,
- A_M — Monotonic h-dim lattice,
- A_U — Unimodal h-dim lattice.

Theorem (Overfitting probability A_B, A_M, and A_U.)

$$P_{\mathbb{F}}(\varepsilon, A_B) = \sum_{p=0}^{D} \sum_{S=p}^{hD} \sum_{F=0}^{h} \frac{|\omega_p| R^p_{D,h}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

$$P_{\mathbb{F}}(\varepsilon, A_M) = \sum_{\vec{\lambda} \in Y^{h,D}_*} \sum_{\substack{\vec{t} \geq \vec{\lambda}, \\ \|\vec{t}\| \leq D}} \frac{|S_h \vec{\lambda}|}{T(\vec{t})} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

$$P_{\mathbb{F}}(\varepsilon, A_U) = \sum_{\vec{\lambda} \in Y^{h,D}_*} \sum_{\substack{\vec{t} \geq \vec{\lambda}, \\ \|\vec{t}\| \leq D}} \sum_{\substack{\vec{t'} \geq \vec{0}, \\ \|\vec{t'}\| \leq D}} \frac{|S_h \vec{\lambda}| \cdot 2^{n(\vec{\lambda})}}{T(\vec{t} + \vec{t'})} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s_0),$$

where $H^{\ell',m}_{L'}(s_0)$ — hypergeometric distribution.

Сравнение сеток и связки монотонных цепочек

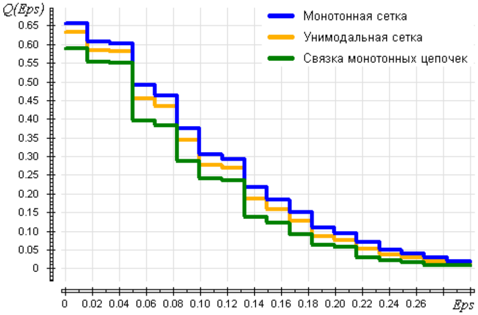


Рис.: Сравнение при разных ε; $D = 5$, $m = 5$, $L = 50$, $\ell = 30$.

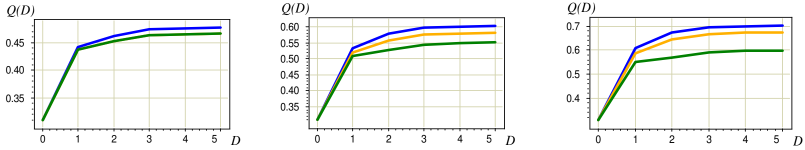
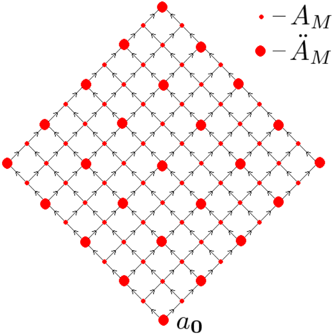


Рис.: Сравнение при разных D, в размерностях $H = 1(2)$, $H = 2(4)$ и $H = 3(6)$. $\varepsilon = 0.04$, $m = 5$, $L = 50$, $\ell = 30$.

Разреженная монотонная сетка



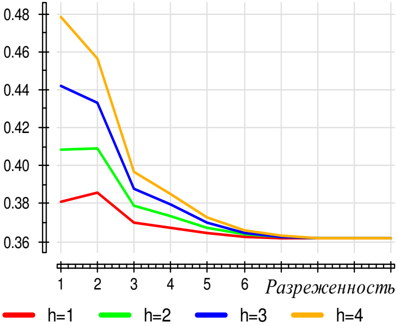


Рис.: Узлы сетки соответствуют алгоритмам, направление стрелок — возрастанию числа ошибок алгоритмов.

Рис.: Зависимость $Q_\mu(\varepsilon, \ddot{A}_M)$ от разреженности монотонной сетки при $L = 100$, $\ell = 60$, $\varepsilon = 0.04$, $D = 12$, $m = 5$.

Результаты и выводы

- Предложен теоретико-групповой подход для вывода формул вероятности переобучения;
- Получены теоретические результаты для несвязного множества алгоритмов;
- Предложено два семейства для аппроксимации сеток их подмножествами малой мощности:
    - Связки монотонных цепочек;
    - Разреженные сети алгоритмов;
- Экспериментально показано, что точность предложенных аппроксимаций падает с возрастанием размерности и разреженности.